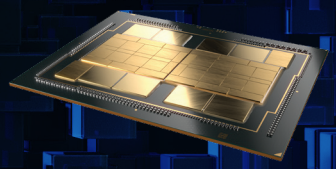
Product Brief

Accelerated Computing
Systems and Graphics

Intel® Data Center GPU Max Series



Maximize impact with the Intel® Data Center GPU Max Series, Intel's highest performing, highest density discrete GPU, which packs more than 100 billion transistors into a package and contains up to 128 Xe cores, Intel's foundational GPU compute building block.

When deploying GPUs in a high-performance computing (HPC) environment, customers face substantial obstacles and inefficiencies caused by the need to port and refactor code. Their efforts are further hampered by proprietary GPU programming environments that prohibit portability between GPU vendors and often result in inconsistency between CPU and GPU implementations. The need for GPU-level memory bandwidth, at scale, and sharing code investments between CPUs and GPUs for running a majority of the workloads in a highly parallelized environment has become essential.

Intel Data Center GPU Max Series is designed for breakthrough performance in data-intensive computing models used in AI and HPC. Based on the Xe HPC architecture that uses both EMIB 2.5D and Foveros packaging technologies to combine 47 active tiles onto a single GPU, fabricated on five different process nodes, Intel Max Series GPUs enable greater flexibility and modularity in the construction of the SOC.

Intel's foundational GPU compute building block features:

- Up to **408 MB of L2 cache** based on discrete SRAM technology, 64 MB of L1 cache and up to **128 GB** of high-bandwidth memory.
- Up to **128 ray tracing units** built into each Max Series GPU for accelerating scientific visualization and animation.
- AI-boosting **Intel® Xe Matrix Extensions** (XMX) with deep systolic arrays enabling vector and matrix capabilities in a single device.
- **oneAPI** standards-based, multiarchitecture programming and tools, which boost performance and productivity and overcome proprietary programming model lock-in.
- Strong performance highlighted by:
  - **Up to 12.8x** performance gain over 3rd Gen Intel® Xeon® processors on LAMMPS (large-scale atomic/molecular massively parallel simulator) workloads running on Xeon Max CPU with kernels offloaded to six Max Series GPUs and optimized by Intel oneAPI tools.[2]

UP TO **2x** performance
gains over competition on AI and HPC workloads due to large L2 Cache[1]

Solving the world's most challenging problems...faster

Increased density and compute power is helping researchers solve problems currently out of reach – for example, creating a 3D map of a mouse brain, or modeling patient-specific blood flow to determine where to insert a heart stent.

The U.S. Department of Energy's Aurora Supercomputer at Argonne National Laboratory (ANL) is expected to be one of the industry's first supercomputers to feature over 1 exaflop of sustained double-precision performance and over 2 exaflops of peak double-precision performance. Aurora will also be the first to showcase the power of pairing Max Series GPUs and CPUs in a single system, with more than 10,000 blades, each containing six Max Series GPUs and two Xeon Max CPUs.

Accelerating HPC and AI Workloads Across Multiple Architectures

AI models continuously require larger data sets for more effective training. The faster you can process the data, the faster you can train and deploy the model. The GPU accelerates end-to-end AI and data analytics pipelines with libraries optimized for Intel architectures and configurations tuned for HPC and AI workloads, high-capacity storage and high-bandwidth memory.

oneAPI

The entire Intel Max Series product family is unified by oneAPI for a common, open, standards-based programming model to unleash productivity and performance. Intel oneAPI tools include advanced compilers, libraries, profilers and code migration tools to easily migrate CUDA code to open C++ with SYCL. Using oneAPI-optimized deep learning frameworks and machine learning libraries, developers can realize drop-in acceleration for data analytics and machine learning workflows.

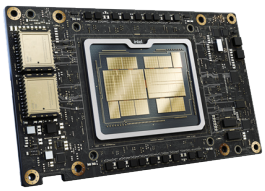
This easy-to-deploy, open-standards approach reduces development time, complexity and cost, and enables developers to overcome the constraints of proprietary environments that limit code portability.

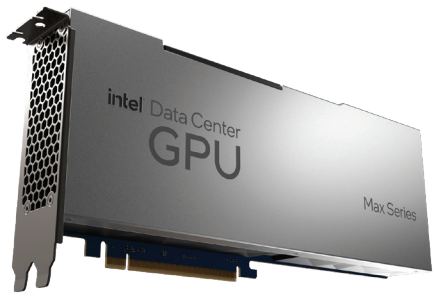
For the latest HPC and AI software developer tools, visit Software for Intel Data Center GPU Max Series.

Intel Data Center Max Series Products & Form Factor Flexibility

Intel Max Series GPUs are available in several form factors:

- **Intel® Data Center Max 1100 GPU**: A 300-watt double-wide AIC card with 56 Xᵉ cores and 48 GB of HBM2E memory. Multiple cards can be connected via Intel Xᵉ Link bridges.
- **Intel® Data Center Max 1450 GPU**: A 600-watt OAM module with 128 Xe cores and 128 GB of HBM that is PRC import friendly. Xelink ports operate at 26.5 GB/s.
- **Intel® Data Center Max 1550 GPU**: Intel's maximum performance 600-watt OAM module with 128 Xᵉ cores and 128 GB of HBM.
- **Intel® Data Center Max Subsystem** with x4 GPU OAM carrier board and Intel Xᵉ Link to enable multi-GPU communication within the subsystem.





Intel Data Center GPU Max Series AIC Card

Intel Data Center GPU Max Series

	Max 1550 GPU (600W OAM)	Max 1450 GPU (600W OAM)	Max 1100 GPU (300W AIC)
Architecture	Xᵉ HPC		
Xᵉ Cores	128	128	56
Memory	HBM2E 128 GB	HBM2E 128 GB	HBM2E 48 GB
Cache	L1 64 MB L2 408 MB	L1 64 MB L2 408 MB	L1 28 MB L2 108 MB
Max TDP	600W	600W	300W
Form Factor	OAM		AIC
Host Interconnect	PCIe Gen5		
Physical Ports	Xᵉ Link 53 GB/s 16 ports	Xᵉ Link 26.5 GB/s 16 ports	Xᵉ Link 53 GB/s 6 ports

For the most up-to-date information, visit Intel.com/MaxSeriesGPU

[1] Visit the Supercomputing 22 page at intel.com/performanceindex for workloads and configurations.

[2] LAMMPS (Atomic Fluid, Copper, DPD, Liquid_crystal, Polyethylene, Protein, Stillinger-Weber, Tersoff, Water)

- Intel® Xeon® 8380: Test by Intel as of 10/11/2022. 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, NUMA configuration SNC2, Total Memory 256 GB (16x16GB 3200 MT/s, Dual-Rank), BIOS Version SE5C620.86B.01.01.0006.2207150335, ucode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, impi:2021.6.0, tbb:2021.6.0; threads/core:; Turbo:on; BuildKnobs: -O3 -ip -xCORE-AVX512 -g -debug inline-debug-info -qopt-zmm-usage=high;
- 4th Gen Intel® Xeon® Scalable Processor: Test by Intel as of 9/29/2022. 1-node, 2x Intel® Xeon® 8480+, HT On, Turbo On, SNC4, Total Memory 512 GB (16x32GB 4800 MT/s, DDR5), BIOS Version SE5C7411.86B.8713.D03.2209091345, ucode revision=0x2b000070, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, impi:2021.6.0, tbb:2021.6.0; threads/core:; Turbo:off; BuildKnobs: -O3 -ip -xCORE-AVX512 -g -debug inline-debug-info -qopt-zmm-usage=high;
- Intel® Xeon® CPU Max Series: Test by Intel as of 9/29/2022. 1-node, 2x Intel® Xeon® Max 9480, HT ON, Turbo ON, NUMA configuration SNC4, Total Memory 128 GB (HBM2E at 3200 MHz), BIOS Version SE5C7411.86B.8424.D03.2208100444, ucode revision=0x2c000020, CentOS Stream 8, Linux version 5.19.0-rc6.0712.intel_next.1.x86_64+server, LAMMPS v2021-09-29 cmkl:2022.1.0, icc:2021.6.0, impi:2021.6.0, tbb:2021.6.0; threads/core:; Turbo:off; BuildKnobs: -O3 -ip -xCORE-AVX512 -g -debug inline-debug-info -qopt-zmm-usage=high;

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Intel technologies may require enabled hardware, software or service activation.